带有超长方体约束的少数类样本生成机制 *

贺作伟 [1]，陶佳晴 [1]，冷强奎 [2†]，翟军昌 [1]，孟祥福 [2]

(1. 渤海大学 信息科学与技术学院, 辽宁 锦州 121013; 2. 辽宁工程技术大学 电子与信息工程学院, 辽宁 葫芦岛 125105)

摘　要：合成少数类过采样技术(SMOTE)是解决类不平衡问题的有效方法之一。但是，SMOTE 的线性插值机制将合成样本限制在原始样本的连线上，导致新样本缺乏多样性，并且这条连线穿过多数类区域时可能会生成噪声样本。针对上述问题，提出一种带有超长方体约束的少数类样本生成机制。该机制使用超长方体作为新样本的生成区域来代替线性插值，以增加合成样本与原始样本的差异性。并通过检测超长方体内是否存在多数类样本来决定是否修正此超长方体，从而防止新合成样本落入多数类区域内。使用所提机制替换线性插值，并集成在三种过采样方法 SMOTE、Borderline-SMOTE 和 ADASYN 中，然后在 KEEL 的 11 个标准数据集上进行了实验评估。结果表明，相比于原始方法，集成后的方法能够帮助分类器取得更高的 F1 值和相当的 G-mean。这说明超长方体生成机制能够显著改善分类器对少数类样本的识别能力，并且能够兼顾到多数类样本。

关键词：不平衡分类；过采样技术；SMOTE；生成机制；超长方体约束

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2022.03.0099

Generation mechanism for minority samples with hypercuboid constraints

He Zuowei[1], Tao Jiaqing[1], Leng Qiangkui[2†], Zhai Junchang[1], Meng Xiangfu[2]

(1. *College of Information Science & Technology*, *Bohai University*, *Jinzhou Liaoning* 121013, *China*; 2. *School of Electronics & Information Engineering*, *Liaoning Technical University*, *Huludao Liaoning* 125105, *China*)

Abstract: Synthetic minority oversampling technology (SMOTE) is one of the effective methods to solve the class-imbalanced problem. However, the linear interpolation mechanism of SMOTE restricts the synthesized samples to the connecting line of the original samples, resulting in a lack of diversity for new samples, and may generate noisy samples when this line passes through the majority class region. In response to the above issues, this paper proposed a generation mechanism for minority samples with hypercuboid constraints. This mechanism constructed a hypercuboid as the generation region of new samples instead of linear interpolation, thereby increasing the variability between the synthesized samples and the original samples. Then, it detected whether there were majority samples in the hypercuboid to determine whether to adjust the hypercuboid, which aimed at preventing the new samples into the region of the majority class. This paper integrated the proposed mechanism into three oversampling methods, i. e. , SMOTE, Borderline-SMOTE, and ADASYN, by using it to replace linear interpolation, and then experimentally evaluated the integrated methods on 11 benchmark datasets from KEEL. The results showed that compared to the original methods, the integrated methods could help the classifier to obtain higher F1 and comparable G-mean. It verifies that the hypercuboid generation mechanism can significantly improve the classifier's ability to recognize minority samples, and meanwhile the majority samples are also taken into account.

Key words: imbalanced classification; oversampling technique; SMOTE; generation mechanism; hypercuboid constraints

0　引言

针对不平衡数据的分类问题是机器学习与数据挖掘领域面临的一项挑战[1, 2]。在二分类问题中，数据不平衡意味着少数类样本的数量远远小于多数类样本的数量[3, 4]。这种类间不平衡会引起标准分类器的偏斜，即分类面更容易被推向少数类样本，导致部分少数类样本不能被正确识别[5]。然而，在一些重要的应用领域中，如医学诊断[6]、软件缺陷预测[7]、恶性肿瘤分级[8]等，少数类通常包含更关键的信息[9]。因此，如何提高针对少数类样本的分类性能是不平衡学习中的关键问题。

目前，处理数据不平衡问题的方法可以分为两类[10]：算法层面方法和数据层面方法。算法层面方法通过修改分类器来强调其对少数类样本的重视[11, 12]。数据层面方法在分类器介入之前先对输入样本进行预处理，以减少数据不平衡的影响[13, 14]。数据层面方法主要包括欠采样技术和过采样技术。欠采样技术通过移除部分多数类样本来实现平衡，但是可能丢失重要的分布信息[15, 16]。而过采样通过增加少数类样本使数据集达到平衡，其中最经典的方法是 Chawla 等[17]提出的合成少数类过采样技术(Synthetic Minority Oversampling TEchnique，SMOTE)。SMOTE 通过在原始的少数类样本之间进行线性插值[18]来生成新的少数类样本，能够提高分类器在测试集上的泛化能力。

近年来，许多 SMOTE 类的方法被相继提出。这些方法或关注类间不平衡问题或致力于改善类内不平衡问题[19]。对

收稿日期：2022-03-19；**修回日期：**2022-05-05　**基金项目：**国家自然科学基金资助项目(61602056、61772249)；辽宁省自然科学基金资助项目(2019-ZD-0493)；辽宁省教育厅科研项目(LQ2019012)

作者简介：贺作伟(1997-)，男，山东济宁人，硕士研究生，主要研究方向为人工智能、机器学习；陶佳晴(1998-)，女，辽宁沈阳人，硕士研究生，主要研究方向为人工智能、机器学习；冷强奎(1981-)，男(通信作者)，辽宁建平人，教授，博导，博士，主要研究方向为人工智能与机器学习(qkleng@126.com)；翟军昌(1978-)，男，辽宁丹东人，副教授，硕导，博士，主要研究方向为智能优化算法及其应用；孟祥福(1981-)，男，辽宁朝阳人，教授，博导，博士，主要研究方向为大数据分析及应用.

于类间不平衡问题，Han 等[20]认为位于类边界的少数类样本更容易被误分类，并提出一种只针对边界少数类样本进行合成过采样的 Borderline-SMOTE 方法。He 等[21]提出一种自适应合成少数类过采样技术 ADASYN，该技术根据近邻中多数类样本所占比例来决定少数类样本的合成权重。但无论是 Borderline-SMOTE 还是 ADASYN，它们受近邻参数 K 的影响很大，当 K 取不同的值时，新合成样本的分布具有明显的差异。严等[22]提出基于构造性覆盖算法的过采样技术 CMOTE，该技术根据覆盖密度进行根样本的选择。但对于两个阈值参数 P 和 D 的设定，一直是需要探讨的问题。王等[23]提出的 AdaN_SMOTE 根据精度下降来自适应地决定少数类的近邻值，并根据噪声等其他因素调整近邻大小。该方法合成的新样本能够保留少数类样本明显的聚类特征，并可以有效避免噪声、小分离和复杂形状的影响。李等[24]通过融合支持度 SD 和影响因素 posFac 来指导边界样本的合成，它不仅可以避免 SMOTE 方法选择样本的盲目性，而且能够综合考虑总体样本的分布情况。但 SDRSMOTE 算法仍需要进一步优化，以提高其运行效率。

对于类内不平衡问题(指少数类样本呈多聚簇分布[25])，盛等[26]使用 Box-Cox 变换和 σ 准则改进了密度峰值聚类，并将其与 SMOTE 算法相结合。该方法能够有效剔除各类噪声数据，且获得的聚簇不受空间形状限制，避免了手动输入参数带来的主观因素干扰。Bunkhumpornpat 等[27]将少数类划分为多个任意形状的子聚簇，然后在随机选择的少数类样本与子聚簇中心之间合成新样本。 然而，该方法容易导致类间数据发生重叠，且不能有效标识具有较高过采样权重的边界样本。Nekooeimehr 等[28]提出了一种自适应半/无监督加权过采样方法 A-SUMO。在使用层次聚类算法后，它自适应确定每个子聚簇的过采样大小。此外，A-SUMO 在标识边界样本方面也取得了较好的效果。但也要指出，该方法在聚类时仅考虑距离因素，忽略了样本分布信息，导致抗噪声干扰能力较弱。Douzas 等[29]提出了一种基于 K-Means 和 SMOTE 的启发式过采样方法。它根据每个聚簇的大小和密度来估计采样权重。然而，K-Means 聚类算法无法找到任何不规则的聚簇。并且该方法也未提供可行的策略来确定最佳聚簇数。Tao 等[30]使用密度峰值聚类算法来改进 K-Means 算法在处理类内不平衡问题上的不足。根据欧式距离和密度分布，少数类样本的合成权重被自适应地计算，边界和低密度样本将获得更高的采样机会。尽管该方法能够有效避免噪声数据的合成，但安全距离阈值的设定依赖于一个待调参数 γ，它的合理取值区间目前只能通过实验来获得。

实际上，每一个 SMOTE 类的方法均可被分解为两个机制：数据选择机制和数据生成机制。而上述这些方法均是对数据选择机制的改进，它们在生成新样本时采用与 SMOTE 相同的线性插值。然而，这种线性插值方式限制了合成样本的数据质量，同时它也是一些过采样方法不能克服类内不平衡问题的主要原因[31]。文献[32]也指出，合成的新样本应该具备扩展少数类区域的能力，以强调少数类在数据总体分布中的重要性。特别地，当少数类样本为多聚簇分布时，线性插值会在聚簇之间执行合成操作，这将导致新样本落入多数类区域而形成噪声，并进一步加重两类数据之间的重叠[33]。

为了解决线性插值生成机制存在的问题，并使新样本更具随机性和多样性，本文提出一种带有超长方体约束的数据生成机制(简称超长方体生成机制)。该机制首先以少数类根样本及其选定近邻的连线为对角线，构造一个超长方体，新样本将在此超长方体内生成。但在生成之前，需要检测此超长方体内是否存在多数类样本，若存在，则修正此超长方体。最后，在没有多数类样本的安全区域内生成新的少数类样本。超长方体生成机制是一个独立模块，能够替换线性插值，并可被集成在多数 SMOTE 类的方法中。接下来，本文将首先解释本文提出的超长方体生成机制，然后将其嵌入到 SMOTE、Borderline-SMOTE 和 ADASYN 三种过采样方法中，并与原始方法进行实验对比，以评估该机制的有效性。

1 超长方体数据生成机制

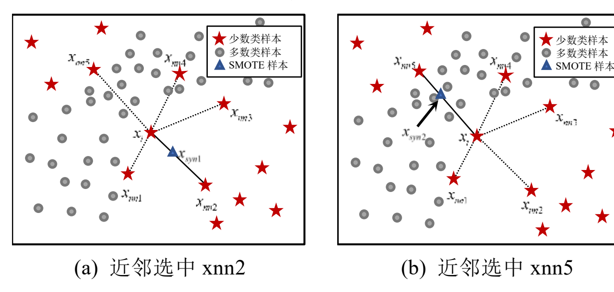
1.1 SMOTE 和线性插值

SMOTE 以迭代搜索方式[34]依次从少数类中选择一个样本作为根样本，并计算根样本到其他少数类样本之间的欧式距离，得到距离根样本最近的 k 个少数类近邻。然后，在根样本和其随机选择的一个近邻之间，使用线性插值合成新的少数类样本。

给定 d 维欧式空间 $\boldsymbol{R}^d$ 中的少数类样本集 $\boldsymbol{X}$，假设 $\boldsymbol{x}_i \in \boldsymbol{X}$ 是当前选定的根样本，在 $k=5$ 时，得到 $\boldsymbol{x}_i$ 的近邻集合 $\boldsymbol{S}=\{\boldsymbol{x}_{nn1},\boldsymbol{x}_{nn2},\boldsymbol{x}_{nn3},\boldsymbol{x}_{nn4},\boldsymbol{x}_{nn5}\}$。根据 SMOTE 的线性插值原理，如果 $\boldsymbol{x}_{nn2}$ 被随机选中(图 1(a))，则新样本 $\boldsymbol{x}_{syn1}$ 将被合成在 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_{nn2}$ 的连线上，即：

$$\boldsymbol{x}_{syn1}=\boldsymbol{x}_i+\varepsilon(\boldsymbol{x}_{nn2}-\boldsymbol{x}_i) \tag{1}$$

其中，ε 是一个(0, 1)之间的随机数。直观来看，$\boldsymbol{x}_{syn1}$ 被限制在一条线段上，文献[31]也指出这种线性插值将影响合成新样本的质量。另外，如果选定的近邻样本为 $\boldsymbol{x}_{nn5}$ (图 1(b))，则 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_{nn5}$ 的连线将穿过多数类区域，新样本 $\boldsymbol{x}_{syn2}$ 将在多数类样本之间合成，从而导致噪声的产生。



(a) 近邻选中 xnn2　　(b) 近邻选中 xnn5

图 1 SMOTE 生成新样本示意图

Fig. 1 Illustration of generating new samples by SMOTE

1.2 超长方体内生成

为了解决上述线性插值存在的问题，本文提出超长方体生成机制来扩展少数类样本的分布范围。给定 $\boldsymbol{R}^d$ 中的少数类根样本 $\boldsymbol{x}_i \in \boldsymbol{X}$，如果它的近邻 $\boldsymbol{x}_{nn2}$ 被随机选中(图 2(a))，则新样本 $\boldsymbol{x}_{syn3}$ 将在 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_{nn2}$ 确定的超长方体内合成，即：

$$\boldsymbol{x}_{syn3}=\boldsymbol{x}_i+\boldsymbol{A}\times(\boldsymbol{x}_{nn2}-\boldsymbol{x}_i) \tag{2}$$

其中，$\boldsymbol{A}=\mathrm{diag}\{\alpha_1,\alpha_2,\ldots,\alpha_d\}$ 是 d 阶对角矩阵，$\alpha_i(i=1,2,\ldots,d)$ 是一个(0, 1)之间的随机数。如果本文将少数类样本按维度展开，则 $\boldsymbol{x}_i=\{x_i^1,x_i^2,\ldots,x_i^d\}^T$，$\boldsymbol{x}_{nn2}=\{x_{nn2}^1,x_{nn2}^2,\ldots,x_{nn2}^d\}^T$，$\boldsymbol{x}_{syn3}$ 将被表示为

$$\boldsymbol{x}_{syn3}=\begin{Bmatrix}(x_i^1-\alpha_1(x_{nn2}^1-x_i^1))\\(x_i^2-\alpha_2(x_{nn2}^2-x_i^2))\\\vdots\\(x_i^d-\alpha_d(x_{nn2}^d-x_i^d))\end{Bmatrix} \tag{3}$$

通过式(3)可以看出，对比于线性插值，在超长方体内生成将增加新样本的随机性和分布范围。但值得注意的是，$\boldsymbol{x}_{syn3}$ 还是存在一定可能被合成在 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_{nn2}$ 的连线上，这时超长方体生成机制就退化为线性插值。

这种退化概率是可以被估计的，假设 α_i 包含 r 位小数，则合成样本位于 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_{nn2}$ 连线上的概率为

$$P=\frac{1}{(10^r)^{d-1}} \tag{4}$$

例如，当 $d=2$ 且 $r=2$ 时，$P=0.01$；当 $d=3$ 且 $r=2$ 时，$P=0.0001$。由此可得到，当高维数据被应用时，超长方体生成机制退化为线性插值的概率是非常低的。

1.3　防止噪声生成策略

如图 2(b)所示，当选定的近邻样本为 $\boldsymbol{x}_{nn5}$ 时，由 $\boldsymbol{x}_{nn5}$ 和 $\boldsymbol{x}_i$ 确定的超长方体与多数类区域发生重叠。如果在该超长方体内合成新样本，则这个新样本会落在多数类样本之间而形成噪声。为了避免合成噪声，本文为超长方体生成机制附加了一个检测及修正策略。首先，计算并检测落入该超长方体内的多数类样本，即得到 $\boldsymbol{y}_{nn1}$ - $\boldsymbol{y}_{nn4}$ 。然后，从 $\boldsymbol{y}_{nn1}$ - $\boldsymbol{y}_{nn4}$ 中找到距离 $\boldsymbol{x}_i$ 最近的多数类样本 $\boldsymbol{y}_{nn4}$ 。最后，执行修正策略，将初始由 $\boldsymbol{x}_{nn5}$ 和 $\boldsymbol{x}_i$ 确定的超长方体修正为由 $\boldsymbol{y}_{nn4}$ 和 $\boldsymbol{x}_i$ 确定的新的超长方体，并最终以其作为新样本的生成区域。

下面给出该检测及修正策略的形式化描述。给定多数类样本集 $\boldsymbol{Y} \in \boldsymbol{R}^d$ ，该策略首先检测 $\boldsymbol{y}_j \in \boldsymbol{Y}$ 是否位于初始超长方体内。对于 $\boldsymbol{y}_j$ 的第 t 维，判断依据如式(5)所示。

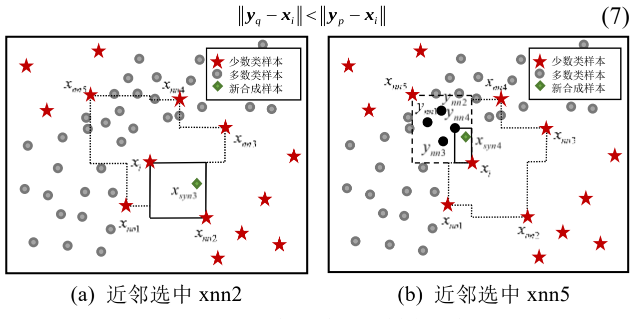
$$\mathrm{Min}(x_i^t, x_{nn5}^t) \le y_j^t \le \mathrm{Max}(x_i^t, x_{nn5}^t) \tag{5}$$

如果 $\boldsymbol{y}_j$ 的每一维度均满足式(5)，则表明 $\boldsymbol{y}_j$ 位于初始超长方体内，此时将 $\boldsymbol{y}_j$ 放入集合 $\boldsymbol{T}$ 中。上述检测步骤要遍历 $\boldsymbol{Y}$ 中每一个样本，遍历完成后若 $\boldsymbol{T} \ne \varnothing$ ，则从 $\boldsymbol{T}$ 中找到距离 $\boldsymbol{x}_i$ 最近的 $\boldsymbol{y}_p$ ：

$$\boldsymbol{y}_p = \mathop{\arg\min}_{\boldsymbol{y}_j} \{\|\boldsymbol{y}_j - \boldsymbol{x}_i\|, \boldsymbol{y}_j \in \boldsymbol{T}\} \tag{6}$$

然后，使用修正策略依据 $\boldsymbol{y}_p$ 和 $\boldsymbol{x}_i$ 重新构造超长方体，新样本将在修正后的超长方体内生成。需要说明的是，修正策略只需执行一次即可保证修正后的超长方体中不包含多数类样本。这是因为，如果存在 $\boldsymbol{y}_q$ ($\boldsymbol{y}_q \in \boldsymbol{T}$ 且 $\boldsymbol{y}_q \ne \boldsymbol{y}_p$)落入修正后的超长方体内，则式(7)成立，这显然与式(6)矛盾。

$$\|\boldsymbol{y}_q - \boldsymbol{x}_i\| < \|\boldsymbol{y}_p - \boldsymbol{x}_i\| \tag{7}$$



(a) 近邻选中 xnn2　　(b) 近邻选中 xnn5

图 2　超长方体机制生成新样本示意图

Fig. 2　Illustration of generating new samples by hypercuboid mechanism

1.4　算法描述

超长方体生成机制的操作步骤如算法 1 所示。其中，第 4-8 步用于检测某个多数类样本 $\boldsymbol{y}_j$ 是否位于由 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_{nn}$ 构造的初始超长方体内。若是，将 $\boldsymbol{y}_j$ 放入集合 $\boldsymbol{T}$ 中。第 13-15 步用于找到 $\boldsymbol{T}$ (不为空时)中距离 $\boldsymbol{x}_i$ 最近的样本 $\boldsymbol{y}_p$ 。第 16-18 步用于合成新的少数类样本。在具体细节上，$|\boldsymbol{Y}|$表示集合 $\boldsymbol{Y}$ 的基数，$flag$ 用作 $\boldsymbol{y}_j$ 是否存入 $\boldsymbol{T}$ 的开关。

该算法的时间复杂度可被估计为 $O(d|\boldsymbol{Y}|)$ ，要高于线性插值的 $O(d)$ 。但由于该机制需嵌入到合成过采样算法中，而过采样过程属于数据预处理阶段，是独立于分类器的，因此它不会对分类器的训练时间构成影响。

算法 1　超长方体数据生成机制

输入：少数类根样本 $x_i = \{x_i^1, x_i^2, \dots, x_i^d\}$ ，近邻 $x_{nn} = \{x_{nn}^1, x_{nn}^2, \dots, x_{nn}^d\}$ ，多数类样本集合 Y 。

输出：一个合成的少数类样本 x_{syn} 。

a)　初始化 $T = \varnothing$;
b)　For　$j = 1$　to　$|Y|$
c)　　$flag = 1$;
d)　　For　$t = 1$　to　d
e)　　　If　$y_j^t > \mathrm{Max}(x_i^t, x_{nn}^t)$　or　$y_j^t < \mathrm{Min}(x_i^t, x_{nn}^t)$　then
f)　　　　$flag = 0$; goto step i);
g)　　　End If
h)　　End For
i)　　If　$flag == 1$　then
j)　　　$T \leftarrow y_j$;
k)　　End If
l)　End For
m)　If　$T \ne \varnothing$　then
n)　　$y_p = \mathop{\arg\min}_{y_j} \{\|y_j - x_i\|, y_j \in T\}$; $x_{nn} = y_p$;
o)　End If
p)　For　$t = 1$　to　d
q)　　$x_{syn}^t = x_i^t + random(0,1) * (x_{nn}^t - x_i^t)$;
r)　End For

需要说明的是，如果本文不把超长方体生成机制当作一个独立模块并用于替代 SMOTE 类过采样算法中的线性插值，那么算法 1 中的修正查找过程可以进一步优化。本文可以预先计算得到训练集中任意两个少数类样本所构成超长方体中包含的多数类样本的信息，然后在每次合成新样本时利用此信息。

令 $G(i,nn)$ 表示由少数类样本 x_i 和 x_{nn} 所构成超长方体中包含的多数类样本的索引。例如， $G(1,5) = \{2,3,7,9\}$ 表明由 x_1 和 x_5 构成的超长方体中包含多数类样本 y_2, y_3, y_7, y_9 。在过采样之前，本文将所有的 $G(i,nn)$ 均计算出来，那么在合成新样本时就可以直接使用这些信息，这将大大缩短算法的运行时间。值得注意的是， $G(i,nn) = G(nn,i)$ 。

在获得 $G(i,nn)$ 的基础上，本文没有必要再遍历整个多数类样本集 Y 。相应地，算法 1 中步骤 b)~l)可以简化为一个步骤，即 $T \leftarrow y_j, j \in G(i,nn)$ 。此时算法 1 的输入中需要包含一个新的参数 $G(i,nn)$ ，算法的时间复杂度将由 $O(d|\boldsymbol{Y}|)$ 下降到 $O(d|G(i,nn)|)$ 。

2　实验结果与分析

提出的超长方体生成机制是一个独立模块，可被嵌入到 SMOTE 类算法中替换线性插值以改善合成数据的质量。本文将所提机制嵌入到 SMOTE、Borderline-SMOTE(简称为 BLSMOTE)、ADASYN 三个过采样算法中， 嵌入后的算法称为 HC-SMOTE、HC-BLSMOTE、HC-ADASYN，然后分别通过人工合成数据集实验和标准数据集实验来评估该机制的有效性。

2.1　人工合成数据集实验

人工合成数据集如图 3 所示，其中少数类样本用红色星形表示，多数类样本用灰色圆形表示。图 3(a)(c)(e)分别表示使用原始的 SMOTE、BLSMOTE、ADASYN 对少数类样本进行过采样后的结果，新合成样本使用三角形表示；图 3(b)(d)(f)分别表示使用 HC-SMOTE、HC-BLSMOTE、HC-ADASYN 进行过采样的结果，新合成样本使用菱形表示。

从图 3 可以看出，SMOTE、BLSMOTE、ADASYN 使用线性插值方式合成少数类样本，新样本均位于原始少数类样本之间的连线上，呈现出明显的线段分布；嵌入超长方体生成机制后，HC-SMOTE、HC-BLSMOTE、HC-ADASYN 合成了分布更为均匀的少数类样本，并且扩展了少数类的分布范围。另外，图 3(a)(c)出现了合成样本跨越多数类区域的情况，这些新样本会成为噪声而使得分类器性能下降。但使用本文所提机制中的防止噪声生成策略后，这种情况不再发生，如图 3(b)(d)所示。

2.2　标准数据集实验

为了体现客观性，从 KEEL 不平衡数据库[35]中选择 11 个标准数据集进行实验，数据集描述见表 1。每一个数据集中均已采用 5 折交叉验证方式划分为训练集和测试集，实验结果将报告 5 次实验的平均值。实验参数按默认设置，

SMOTE、BLSMOTE、ADASYN 在合成样本时近邻参数分别为 5、5、7，BLSMOTE 在判定边界样本时近邻参数为 7。分类器使用 C4.5[36]和 AdaBoost[37]。

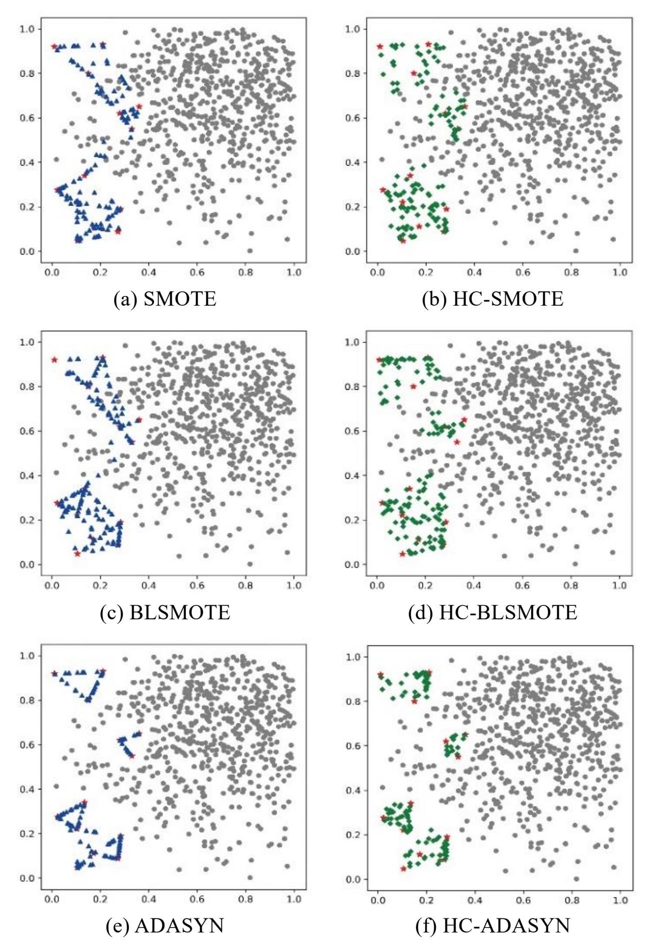


图 3　原始方法(左)与集成后方法(右)的对比

Fig. 3　Comparison of the original methods (left) with the integrated methods (right)

评价指标包括 *F*1 和 *G-mean*。*F*1 是精准率(*Precision*)和召回率(*Recall*)的调和平均，反映分类器对少数类样本的分类能力；*G-mean* 是敏感性(*Sensitivity*)和特异性(*Specificity*)的几何平均，反映分类器兼顾两类样本时的分类能力。这些指标均以混淆矩阵(表 2)为基础，计算公式为

$$Precision = \frac{TP}{FP + TP} \tag{8}$$

$$Recall = Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{11}$$

$$G-mean = \sqrt{Sensitivity \times Specificity} \tag{12}$$

表 1　数据集基本信息

Tab. 1　Description of the datasets

数据集名称	关键字	样本个数	维度	不平衡率
Wisconsin	Wi	683	9	1.86
Vehicle2	Ve2	846	18	2.88
New-thyroid2	N2	215	5	5.14
Segment0	S0	2308	19	6.02
Glass6	G6	214	9	6.38
Yeast3	Y3	1484	8	8.10
Vowel0	Vo0	988	13	9.98
Ecoli4	E4	336	7	15.80
Page-blocks-1-3_vs_4	P4	472	10	15.86
Shuttle-c2-vs-c4	S4	129	9	20.50
Yeast5	Y5	1484	8	32.73

表 2　二分类问题的混淆矩阵

Tab. 2　Confusion matrix of the two-class problem

实际情况	预测为少数类	预测为多数类
实际为少数类	*TP*	*FN*
实际为多数类	*FP*	*TN*

表 3 给出了 HC-SMOTE 与原始 SMOTE 的对比实验结果。C4.5 分别在 9 个数据集上和 5 个数据集上取得了更高的 *F*1 和 *G-mean*，这说明经过 HC-SMOTE 过采样后，显著提升了 C4.5 针对少数类样本的识别能力，但在兼顾多数类方面还存在不足；AdaBoost 在全部 11 个数据集上均取得了更高的 *F*1，同时也在 8 个数据集上获得了更好的 *G-mean*，这说明 HC-SMOTE 对 AdaBoost 产生了积极影响。

表 3　SMOTE 与 HC-SMOTE 的对比

Tab. 3　Comparison of SMOTE with HC-SMOTE

Data	C4.5				AdaBoost			
	*F*1		*G*-mean		*F*1		*G*-mean	
	SMOTE	HC-SMOTE	SMOTE	HC-SMOTE	SMOTE	HC-SMOTE	SMOTE	HC-SMOTE
Wi	**0.9460**	0.9278	**0.9600**	0.9470	0.9383	**0.9454**	0.9556	**0.9580**
Ve2	**0.9230**	0.9202	**0.9586**	0.9546	0.9176	**0.9453**	0.9592	**0.9643**
N2	0.7918	**0.9405**	0.9258	**0.9527**	0.8695	**0.9275**	0.9484	**0.9503**
S0	0.9730	**0.9804**	0.9878	**0.9929**	0.9792	**0.9818**	**0.9952**	0.9919
G6	0.7785	**0.7809**	**0.8888**	0.8829	0.7484	**0.8508**	0.8615	**0.9173**
Y3	0.7482	**0.7723**	0.9033	**0.9209**	0.7125	**0.7732**	0.8677	**0.9027**
Vo0	0.8988	**0.9166**	**0.9786**	0.9600	0.9068	**0.9307**	0.9747	**0.9770**
E4	0.7141	**0.7691**	**0.9502**	0.8789	0.7245	**0.7429**	**0.9282**	0.8519
P4	0.8539	**0.9667**	0.9887	**0.9977**	0.8595	**0.9667**	0.9887	**0.9977**
S4	0.9600	**1.0000**	0.9958	**1.0000**	0.9000	**1.0000**	0.9826	**1.0000**
Y5	0.6584	**0.7136**	**0.9401**	0.9193	0.6803	**0.7246**	**0.9273**	0.9200

表 4 给出了 HC-BLSMOTE 与原始 BLSMOTE 的对比实验结果。C4.5 分别在 7 个数据集上和 9 个数据集上取得了更高的 *F*1 和 *G-mean*，AdaBoost 在 8 个数据集上取得了更高的 *F*1 和 *G-mean*。由于 BLSMOTE 在数据选择阶段只对边界少数类样本进行过采样，而本文所提机制与 BLSMOTE 结合后表现出优异的性能，说明在边界处对少数类样本进行超长方体区域内的合成，能够极大改善新合成样本的质量，并有助于提高分类器的泛化性能。

表 5 给出了 HC-ADASYN 与原始 ADASYN 的对比实验结果。C4.5 和 AdaBoost 分别在 11 个和 10 个数据集上取得了更高的 F1，但仅在 6 个和 3 个数据集上取得了更高的 G-mean。ADASYN 为每个少数类样本施加一个合成权重，即当邻域内多数类样本越多时该合成权重越大。在嵌入超长方体生成机制后，HC-ADASYN 将更关注权重大的少数类样本，但可能导致部分多数类样本被忽视。

图 4 是上述实验结果的箱线图，红色菱形点表示平均值，绿色虚线表示中位数。SM、BD、AD 分别是过采样方法 SMOTE、BLSMOTE 和 ADASYN 的缩写。C45 和 Ada 分别是分类器 C4.5 和 AdaBoost 的缩写。从子图 4(a)(c)(e)可以看出，改进后的方法在 *F1* 上取得了大幅的领先，这说明本文所提机制能够明显提升分类器对少数类的识别。同时，改进后的 HC-SMOTE 和 HC-BLSMOTE 在 *G-mean* 的表现上也优于原始方法。整体来看，超长方体生成机制嵌入到 Borderline-SMOTE 后的性能最好。

表 4　BLSMOTE 与 HC-BLSMOTE 的对比

Tab. 4　Comparison of BLSMOTE with HC-BLSMOTE

Data	C4.5				AdaBoost			
	F1		*G*-mean		*F*1		*G*-mean	
	BLSMOTE	HC-BLSMOTE	BLSMOTE	HC-BLSMOTE	BLSMOTE	HC-BLSMOTE	BLSMOTE	HC-BLSMOTE
Wi	0.9243	**0.9414**	0.9456	**0.9604**	0.9244	**0.9442**	0.9417	**0.9590**
Ve2	0.9192	**0.9323**	0.9447	**0.9581**	**0.9566**	0.9560	**0.9728**	0.9681
N2	0.9119	**0.9273**	0.9353	**0.9501**	0.8894	**0.9581**	0.9396	**0.9798**
S0	**0.9759**	0.9717	0.9870	**0.9888**	0.9743	**0.9759**	0.9868	**0.9871**
G6	0.7578	**0.7921**	0.8642	**0.8913**	0.8264	**0.8636**	0.8794	**0.9201**
Y3	0.7773	**0.7939**	0.9117	**0.9274**	0.7423	**0.7454**	0.8638	**0.8793**
Vo0	**0.9381**	0.8544	**0.9882**	0.9425	**0.9576**	0.8996	**0.9905**	0.9638
E4	0.7254	**0.7667**	0.8759	**0.9077**	0.7048	**0.7476**	0.8496	**0.8775**
P4	0.9227	**0.9667**	0.9943	**0.9977**	0.9227	**0.9667**	0.9943	**0.9977**
S4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Y5	**0.7413**	0.7104	0.9331	**0.9420**	0.7190	**0.7371**	0.8722	**0.9429**

表 5　ADASYN 与 HC-ADASYN 的对比

Tab. 5　Comparison of ADASYN with HC-ADASYN

Data	C4.5				AdaBoost			
	F1		*G*-mean		*F*1		*G*-mean	
	ADASYN	HC-ADASYN	ADASYN	HC-ADASYN	ADASYN	HC-ADASYN	ADASYN	HC-ADASYN
Wi	0.9337	**0.9360**	**0.9557**	0.9535	**0.9430**	0.9383	**0.9606**	0.9494
Ve2	0.8977	**0.9334**	0.9446	**0.9555**	0.9229	**0.9489**	0.9601	**0.9641**
N2	0.8211	**0.9405**	0.9456	**0.9527**	0.8772	**0.9275**	**0.9601**	0.9503
S0	0.9595	**0.9773**	0.9891	**0.9898**	0.9677	**0.9833**	0.9893	**0.9908**
G6	0.6888	**0.7448**	**0.8591**	0.8326	0.7458	**0.8407**	**0.8937**	0.8827
Y3	0.7234	**0.7763**	0.8951	**0.9030**	0.7069	**0.7425**	**0.8749**	0.8727
Vo0	0.8521	**0.9125**	**0.9820**	0.9589	0.8970	**0.9363**	**0.9882**	0.9667
E4	0.6134	**0.7076**	**0.8417**	0.8222	0.6181	**0.6333**	**0.8296**	0.7860
P4	0.8300	**0.9227**	0.9864	**0.9943**	0.8803	**0.8833**	**0.9910**	0.9560
S4	0.8476	**1.0000**	0.9830	**1.0000**	0.9600	**1.0000**	0.9958	**1.0000**
Y5	0.6087	**0.7193**	**0.9467**	0.9081	0.7180	**0.7680**	**0.9757**	0.8858

(a) HC-SM 和 SM 的 F1　(b) HC-SM 和 SM 的 G-mean　(c) HC-BD 和 BD 的 F1　(d) HC-BD 和 BD 的 G-mean

(e) HC-AD 和 AD 的 F1　(f) HC-AD 和 AD 的 G-mean

图 4　标准数据集实验结果的箱线图

Fig. 4　Boxplots of experimental results on benchmark datasets

3 结束语

本文提出一种新的数据生成机制来改进合成过采样方法，它使用超长方体作为新样本的生成区域来代替线性插值，以增加新合成样本与原始样本的差异性。为防止新样本落入多数类区域内，一个检测及修正策略被附加到该超长方体生成机制中，从而避免了噪声的产生。

在标准数据集上的实验表明，当该机制集成到SMOTE、Borderline-SMOTE、ADASYN 三个过采样方法后，两个标准分类器在大部分数据集上均取得了更高的 *F*1 值，说明超长方体生成机制能够显著改善分类器对少数类样本的识别能力。在 *G-mean* 评价指标上，集成后的方法表现与原始方法相当，说明其在关注少数类样本的同时，也能够兼顾多数类样本。

本文工作从数据生成机制出发，为不平衡学习中过采样方法的研究提供了一个新的思路。但提出的超长方体生成机制是启发式的，其有效性建立在实验评估的基础之上。下一步工作将在理论层面上深入研究数据生成机制对合成样本质量的影响。

参考文献：

[1] Shi Hongbo, Gao Qigang, Ji Suqin, *et al.* A hybrid sampling method based on safe screening for imbalanced datasets with sparse structure [C]// 2018 International Joint Conference on Neural Networks (IJCNN) . New York: IEEE Press, 2018: 1-8

[2] Li Junnan, Zhu Qingsheng, Wu Quanwang, *et al.* A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors [J]. Information Sciences, 2021, 565: 438-455.

[3] 杨浩，陈红梅. 结合样本局部密度的非平衡数据集成分类算法 [J]. 计算机科学与探索, 2020, 14 (2): 274-284. (Yang Hao, Chen Hongmei. Ensemble classification algorithm for imbalanced data combined with local area density [J]. Journal of Frontiers of Computer Science and Technology, 2020, 14 (2): 274-284.)

[4] Barua S, Islam M M, Yao X, *et al.* MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 26 (2): 405-425.

[5] Zheng Ming, Li Tong, Zheng Xiaoyao, *et al.* UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification [J]. Information Sciences, 2021, 576: 658-680.

[6] Parvin H, Minaei-Bidgoli B, Alizadeh H. Detection of cancer patients using an innovative method for learning at imbalanced datasets [C]// International Conference on Rough Sets and Knowledge Technology. Berlin: Springer Press, 2011: 376-381.

[7] Wang Shuo, Yao Xin. Using class imbalance learning for software defect prediction [J]. IEEE Trans on Reliability, 2013, 62 (2): 434-443.

[8] Krawczyk B, Galar M, Jelen L, *et al.* Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy [J]. Applied Soft Computing, 2016, 38: 714-726.

[9] 徐玲玲，迟冬祥. 面向不平衡数据集的机器学习分类策略 [J]. 计算机工程与应用, 2020, 56 (24): 12-27. (Xyu Lingling, Chi Dongxiang. Machine Learning Classification Strategy for Imbalanced Data Sets [J]. Computer Engineering and Applications, 2020, 56 (24): 12-27.)

[10] He Haibo, Garcia E A. Learning from imbalanced data [J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21 (9): 1263-1284.

[11] Dubey H, Pudi V. Class based weighted k-nearest neighbor over imbalance dataset [C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer Press, 2013: 305-316.

[12] Fan Wei, Stolfo S J, Zhang Junxin, *et al.* AdaCost: misclassification cost-sensitive boosting [C]// Proceedings of the 16th International Conference on Machine Learning (ICML) . San Francisco: Morgan Kaufmann Pub Inc, 1999: 97-105.

[13] Elreedy D, Atiya A F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance [J]. Information Sciences, 2019, 505: 32-64.

[14] Zhu Yuanwei, Yan Yuanting, Zhang Yiwen, *et al.* EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning [J]. Neurocomputing, 2020, 417: 333-346.

[15] Fernández A, Garcia S, Herrera F, *et al.* SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary [J]. Journal of Artificial Intelligence Research, 2018, 61: 863-905.

[16] 吴艺凡，梁吉业，王俊红. 基于混合采样的非平衡数据分类算法 [J]. 计算机科学与探索, 2019, 13 (2): 342-349. (Wu Yifan, Liang Jiye, Wang Junhong. Classification algorithm based on hybrid sampling for unbalanced data [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13 (2): 342-349.)

[17] Chawla N V, Bowyer K W, Hall L O, *et al.* SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[18] Blu T, Thévenaz P, Unser M. Linear interpolation revitalized [J]. IEEE Trans on Image Processing, 2004, 13 (5): 710-719.

[19] Tao Xinmin, Zheng Yujia, Tao Weichen, *et al.* SVDD-based weighted oversampling technique for imbalanced and overlapped dataset learning [J]. Information Sciences, 2022, 588: 13-51.

[20] Han Hui, Wang Wenyuan, Mao Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C]// International Conference on Intelligent Computing. Berlin: Springer Press, 2005: 878-887.

[21] He Haibo, Bai Yang, Garcia E A, *et al.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning [C]// 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) . New York: IEEE Press, 2008: 1322-1328.

[22] 严远亭，朱原玮，吴增宝，等. 构造性覆盖算法的 SMOTE 过采样方法 [J]. 计算机科学与探索, 2020, 14 (6): 975-984. (Yan Yuanting, Zhu Yuanwei, Wu Zengbao, *et al.* Constructive covering algorithm-based SMOTE over-sampling method [J]. Journal of Frontiers of Computer Science and Technology, 2020, 14 (6): 975-984.)

[23] 王芳，吴文通，张立立，等. 邻域自适应 SMOTE 算法研究 [J]. 计算机应用研究, 2021, 38 (06): 1673-1677. (Wang Fang, Wu Wentong, Zhang Lili, *et al.* Research on neighborhood adaptive SMOTE algorithm [J]. Application Research of Computers, 2021, 38 (06): 1673-1677.)

[24] 李克文，林亚林，杨耀忠. 一种改进的基于欧氏距离的 SDRSMOTE 算法 [J]. 计算机工程与科学, 2019, 41 (11): 2063-2070. (Li Kewen, Lin Yalin, Yang Yaozhong. An improved SDRSMOTE algorithm based on Euclidean distance [J]. Computer Engineering & Science, 2019, 41 (11): 2063-2070.)

[25] Leevy J L, Khoshgoftaar T M, Bauder R A, *et al.* A survey on addressing high-class imbalance in big data [J]. Journal of Big Data, 2018, 5 (1): 1-30.

[26] 盛凯，刘忠，周德超，等. 面向不平衡分类的 IDP-SMOTE 重采样算法 [J]. 计算机应用研究, 2019, 36 (01): 115-118. (Sheng Kai, Liu Zhong, Zhou Dechao, *et al.* IDP-SMOTE resampling algorithm for imbalanced classification [J]. Application Research of Computers, 2019, 36 (01): 115-118.)

[27] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. DBSMOTE:

density-based synthetic minority over-sampling technique [J]. Applied Intelligence, 2012, 36 (3): 664-684.

[28] Nekooeimehr I, Lai-Yuen S K. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets [J]. Expert Systems with Applications, 2016, 46: 405-416.

[29] Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE [J]. Information Sciences, 2018, 465: 1-20.

[30] Tao Xinmin, Li Qing, Guo Wenjie, *et al.* Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering [J]. Information Sciences, 2020, 519: 43-73.

[31] Li Yihong, Wang Yunpeng, Li Tao, *et al.* SP-SMOTE: A novel space partitioning based synthetic minority oversampling technique [J]. Knowledge-Based Systems, 2021, 228: 107269.

[32] Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE [J]. Information Sciences, 2019, 501: 118-135.

[33] Zhu Tuanfei, Lin Yaping, Liu Yonghe. Improving interpolation-based oversampling for imbalanced data learning [J]. Knowledge-Based Systems, 2020, 187: 104826.

[34] Raghuwanshi B S, Shukla S. SMOTE based class-specific extreme learning machine for imbalanced learning [J]. Knowledge-Based Systems, 2020, 187: 104814.

[35] Moreno-torres J G, Sáez J A, Herrera F. Study on the impact of partition-induced dataset shift on k-fold cross-validation [J]. IEEE Trans on Neural Networks and Learning Systems, 2012, 23 (8): 1304-1312.

[36] Elyan E, Moreno-garcia C F, Jayne C. CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification [J]. Neural Computing and Applications, 2021, 33 (7): 2839-2851.

[37] Niu Kun, Zhang Zaimei, Liu Yan, *et al.* Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending [J]. Information Sciences, 2020, 536: 120-13.